

## Research



**Cite this article:** Zhou X *et al.* 2016 The Trichoptera barcode initiative: a strategy for generating a species-level Tree of Life. *Phil. Trans. R. Soc. B* **371**: 20160025.  
<http://dx.doi.org/10.1098/rstb.2016.0025>

Accepted: 1 May 2016

One contribution of 16 to a theme issue  
'From DNA barcodes to biomes'.

**Subject Areas:**

evolution, taxonomy and systematics

**Keywords:**

DNA barcodes, caddisfly, phylogeny,  
integrative taxonomy

**Authors for correspondence:**

Xin Zhou

e-mail: [xinzhoucaddis@icloud.com](mailto:xinzhoucaddis@icloud.com)

Karl M. Kjer

e-mail: [kkjer@ucdavis.edu](mailto:kkjer@ucdavis.edu)

<sup>†</sup>Main coordinators of the paper. Other  
co-authors are ordered by last names.

# The Trichoptera barcode initiative: a strategy for generating a species-level Tree of Life

Xin Zhou<sup>1,2,†</sup>, Paul B. Frandsen<sup>3,†</sup>, Ralph W. Holzenthal<sup>4,†</sup>, Clare R. Beet<sup>5</sup>, Kristi R. Bennett<sup>5</sup>, Roger J. Blahnik<sup>4</sup>, Núria Bonada<sup>6</sup>, David Cartwright<sup>7</sup>, Suvdtsetseg Chuluunbat<sup>8</sup>, Graeme V. Cocks<sup>9</sup>, Gemma E. Collins<sup>5</sup>, Jeremy deWaard<sup>10</sup>, John Dean<sup>11</sup>, Oliver S. Flint Jr<sup>12</sup>, Axel Hausmann<sup>13</sup>, Lars Hendrich<sup>13</sup>, Monika Hess<sup>14</sup>, Ian D. Hogg<sup>5</sup>, Boris C. Kondratieff<sup>15</sup>, Hans Malicky<sup>16</sup>, Megan A. Milton<sup>10</sup>, Jérôme Morinière<sup>13</sup>, John C. Morse<sup>17</sup>, François Ngera Mwangi<sup>18</sup>, Steffen U. Pauls<sup>19</sup>, María Razo Gonzalez<sup>20</sup>, Aki Rinne<sup>21</sup>, Jason L. Robinson<sup>22</sup>, Juha Salokannel<sup>23</sup>, Michael Shackleton<sup>24</sup>, Brian Smith<sup>25</sup>, Alexandros Stamatakis<sup>26,27</sup>, Ros StClair<sup>11</sup>, Jessica A. Thomas<sup>28</sup>, Carmen Zamora-Muñoz<sup>29</sup>, Tanja Ziesmann<sup>30</sup> and Karl M. Kjer<sup>31,†</sup>

<sup>1</sup>Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing 100193, People's Republic of China

<sup>2</sup>College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, People's Republic of China

<sup>3</sup>Office of Research Information Services, Office of the Chief Information Officer, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, USA

<sup>4</sup>Department of Entomology, University of Minnesota, 1980 Folwell Avenue, St Paul, MN 55108, USA

<sup>5</sup>School of Science, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

<sup>6</sup>Grup de Recerca Freshwater Ecology and Management (FEM), Departament d'Ecologia, Facultat de Biologia, Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Diagonal, 643, 08028 Barcelona, Catalonia, Spain

<sup>7</sup>13 Brolga Crescent, Wandana Heights, Victoria 3216, Australia

<sup>8</sup>Department of Biology, Mongolian National University of Education, 3rd Palace, Beijing Street, Ulaanbaatar 14191, Mongolia

<sup>9</sup>44 Marks Street, Hermit Park, Queensland 4812, Australia

<sup>10</sup>Centre for Biodiversity Genomics, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada N1G 2W1

<sup>11</sup>Environment Protection Authority Victoria, Ernest Jones Drive, Macleod 3085, Australia

<sup>12</sup>Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012, USA

<sup>13</sup>SNSB-Bavarian State Collection of Zoology, Münchhausenstr. 21, 81247 Munich, Germany

<sup>14</sup>Büro H2-Ökologische Gutachten, Hess+Heckes GbR, Rumpfstraße 42, 80469 München, Germany

<sup>15</sup>Department of Bioagricultural Sciences and Pest Management, Colorado State University, 1177 Campus Delivery, Fort Collins, CO 80523, USA

<sup>16</sup>Biologische Station Lunz, Austrian Academy of Sciences, A-3293 Lunz am see, Austria

<sup>17</sup>Department of Plant and Environmental Sciences, Clemson University, PO Box 340310, Clemson, SC 29634-0310, USA

<sup>18</sup>Centre de Recherche en Sciences Naturelles Lwiro, P.O. D.S Bukavu, D. R. Congo

<sup>19</sup>Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

<sup>20</sup>Unidad Multidisciplinaria de Docencia e Investigación, Universidad Nacional Autónoma de México, Facultad de Ciencias, Campus Juriquilla, Querétaro, 76230, México

<sup>21</sup>Finnish Environment Institute, Merikasarminkatu 8 D, 00160 Helsinki, Finland

<sup>22</sup>Illinois Natural History Survey, Prairie Research Institute at the University of Illinois at Urbana-Champaign, 1816 S. Oak Street, MC 652, Champaign, IL 61820, USA

<sup>23</sup>Aquatic Insects Expert Group of Finland, Siikinkatu 13, 33710, Tampere, Finland

<sup>24</sup>Murray-Darling Freshwater Research Centre, La Trobe University, 133 McKoy Street, Wodonga, Victoria 3691, Australia

<sup>25</sup>National Institute of Water and Atmospheric Research, PO Box, 11115, Hamilton 3240, New Zealand

<sup>26</sup>Scientific Computing Group, Heidelberg Institute for Theoretical Studies (HITS), 69118 Heidelberg, Germany


<sup>27</sup>Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, 35 D-69118 Heidelberg, Germany

<sup>28</sup>BioArch, Environment Building, Department of Biology, University of York, York, YO10 5DD, UK

<sup>29</sup>Department of Zoology, Faculty of Sciences, University of Granada, C/Severo Ochoa s/n, 18071 Granada, Spain

<sup>30</sup>Zoologisches Forschungsmuseum Alexander Koenig (ZFMK)/Zentrum für Molekulare Biodiversitätsforschung (ZMB), Bonn 5 76131 Karlsruhe, Germany

<sup>31</sup>Department of Entomology and Nematology, University of California-Davis, 1282 Academic Surge, Davis, CA 95616, USA

 XZ, 0000-0002-1407-7952; KMK, 0000-0001-7370-9617

DNA barcoding was intended as a means to provide species-level identifications through associating DNA sequences from unknown specimens to those from curated reference specimens. Although barcodes were not designed for phylogenetics, they can be beneficial to the completion of the Tree of Life. The barcode database for Trichoptera is relatively comprehensive, with data from every family, approximately two-thirds of the genera, and one-third of the described species. Most Trichoptera, as with most of life's species, have never been subjected to any formal phylogenetic analysis. Here, we present a phylogeny with over 16 000 unique haplotypes as a working hypothesis that can be updated as our estimates improve. We suggest a strategy of implementing constrained tree searches, which allow larger datasets to dictate the backbone phylogeny, while the barcode data fill out the tips of the tree. We also discuss how this phylogeny could be used to focus taxonomic attention on ambiguous species boundaries and hidden biodiversity. We suggest that systematists continue to differentiate between 'Barcode Index Numbers' (BINs) and 'species' that have been formally described. Each has utility, but they are not synonyms. We highlight examples of integrative taxonomy, using both barcodes and morphology for species description.

This article is part of the themed issue 'From DNA barcodes to biomes'.

## 1. Introduction

Generating an accurate 'Tree of Life' (phylogeny) including every species that exists, and has ever existed, is an impossible challenge. However, systematists work toward this goal, adding parts of the puzzle taxon by taxon. The culture of science includes incentives to work both independently and cooperatively. In contrast with endeavours like the space programme, which cannot be advanced without constant coordination, there is a logical subdivision of labour in systematics, with specialists working, sometimes in isolation, on their own phylogenetically organized taxa, gradually adding consensus and modifying hypotheses, which we hope improve with time as we add both characters and taxa. Our independent work can eventually come together. A sense of urgency has been expressed about the pace of 'completing' the Tree of Life. For example, in the USA, the National Science Foundation's 'Assembling the Tree of Life' and 'Dimensions in Biodiversity' programmes have attempted to accelerate the pace by supporting collaborative efforts with large grants. The challenges are significant. Recently, it was thought that a major bottleneck towards completing the Tree of Life would be in computing

speed, but programs such as RAXML [1] and FastTree [2] permit the rapid and efficient analysis of datasets consisting of thousands of both genes and taxa. The ultimate goal under one way of thinking would be to develop the ability to analyse millions of species simultaneously and summarize them on the same comprehensive tree. A more practical alternative approach would be to subdivide the tree into smaller taxonomic groups, as has been standard practice, with each specialist working on their own group. With this approach, subtrees must be grafted together into a larger Tree of Life. A recent study proposed to align phylogenies from different resources to synthesize a comprehensive Tree of Life, i.e. the Open Tree of Life [3]. We believe, however, that it is critical to identify monophyletic groups carefully *a priori*, to avoid unstable phylogenies for subsequent analysis. The strategy of subdividing the task into smaller subclades is a solution that simply requires some criterion for deciding which clades can be independently analysed.

Systematic biology is facing a radical transition from the standard 'few genes and morphology', PCR-Sanger-based approach, to transcriptomic and genomic sequencing. With new datasets of unprecedented size (e.g. [4–6]), the backbone of the Tree of Life, at least for insects, has been largely resolved. This is a good time to discuss how we might integrate the work on the terminal branches of the tree, such as the DNA barcoding efforts [7]. A balance must be struck between using millions of nucleotides from transcriptomes—representing only dozens of representative taxa—and using barcode data from, for example, 40 000 individuals.

## (a) Using DNA barcodes to build leaves of the Tree of Life

Using DNA sequences to identify specimens has been a possibility since the 1980s. However, initial efforts were uncoordinated and without standard and agreed-upon protocols (i.e. [8,9]). A truly grand vision of using DNA to identify every species on Earth would require a coordination of efforts, with the selection of one, or a few standardized gene fragments, and a huge database of identified sequences and collections of voucher specimens, to which an unknown specimen could be compared and then identified [7,10]. A coordinated international effort (the International Barcode of Life, or iBOL, <http://www.ibol.org/>) to create such a database (Barcode of Life Database, BOLD, <http://www.boldsystems.org/>) has been underway for over 10 years. A 658-nucleotide fragment of the 5'-end of mitochondrial cytochrome oxidase c, subunit 1 (COI) is the most commonly used marker for identifying animal species. This short, standardized DNA fragment is referred to as a 'barcode'.

The BOLD is a powerful tool for organizing, visualizing and downloading DNA sequences, images and collection records, and we have made extensive use of it here. By February 2016, over 6 million specimen records had been registered in BOLD, representing over 250 000 species and approximately 500 000 BINs. In addition to its intended function for species identification, we present a constructive, integrative approach to discovering, describing and understanding biodiversity, using Trichoptera, or caddisflies, as a model taxon. We are particularly interested in Trichoptera taxonomy and phylogenetics. We recognize the limitations of barcode data for generating phylogenies [11,12] and discuss how these limitations might be mitigated for both inferring phylogeny and for discovering and describing biodiversity.

## (b) Using Trichoptera as a model system

Trichoptera are an order of holometabolous insects, with small, moth-like adults and aquatic larvae that produce silk to construct a diverse array of case and retreat architecture [13,14]. The larvae exploit diverse aquatic microhabitats and are important participants in nutrient dynamics and energy flow in freshwater ecosystems. Because different species are differentially sensitive to pollution, their relative diversity and abundance can be used to assess and monitor water quality [15]. It is the larval stage in Trichoptera that is collected for biological monitoring, but because most taxonomy (i.e. species diagnoses and descriptions) is based on male genitalia, many larval species (and females) are difficult to identify and have not been described. We are fortunate in that the monophyly of Trichoptera is well established and its sister taxon relationship with Lepidoptera (moths and butterflies), comprising the super-order Amphiesmenoptera, is the most strongly supported sister order relationship within insects [16]. Therefore, a phylogeny of Trichoptera can be inferred and subsequently rooted with its sister taxon Lepidoptera alone. Then the Amphiesmenoptera phylogeny can be grafted onto a larger phylogeny, derived from an analysis of Holometabola, using representative taxa, rather than all species. Still, generating the phylogeny of Trichoptera, a moderately diverse insect order with over 14 500 described species, is no small task. The most significant challenge is the availability of specimens representing the order's diversity. Most species (and higher taxa) follow a hollow curve distribution [17], referring to the dominance of few abundant species and the presence of many more rare species, some of which may be known as single specimens, collected only once and never seen again. Since 1995, with the help of collaborators, we have collected molecular data from approximately 480 of the 616 extant Trichoptera genera. This includes sampling a leg from every Trichoptera species (under a sampling threshold of 40 years old or less) in both the National Museum of Natural History (Smithsonian Institution) and the University of Minnesota Insect Collection (UMSP). These institutions have employed Trichoptera specialists for 55 and 30 years, respectively. Other specimens have come to us from H.M., who holds one of the largest Trichoptera collections in the world. If one were to combine all the collecting efforts from our previous Trichoptera phylogenetics projects, it would conservatively be represented by over 100 'person years'. Yet over 20% of the genera are missing, many of which will probably remain unobtainable (e.g. *nomina dubia* or rare, monotypic genera). According to the Trichoptera World Checklist [18], as of January 2013 there were 14 548 extant species of Trichoptera, making it the seventh most species-rich of the 30 insect orders. There are probably another 15 000 yet to be described (our minimum best professional estimate). While it may be possible to imagine a 'Tree of Life' for mammals or birds, the scale of the endeavour for most of life's representatives, insects and other arthropods as well as bacteria, is far greater. However, even if our efforts represent only one-eighth of the Trichoptera tree, we believe it is worth putting together a phylogeny from the taxa that we have now. The overwhelming majority of species in our dataset are represented with only the barcode data. Many of these species have never been subjected to any kind of formal phylogenetic analysis. Here, we reflect on what we are doing with Trichoptera, toward reaching our phylogenetic goals, and suggest that our approach may be applicable to other taxa.

Earlier work on a single diverse genus, *Chimarra* [12], showed that many nodes from COI alone were congruent with our best estimates of phylogeny from multiple genes and morphology, and those that were 'unexpected' were weakly supported and easy to identify. Closely related species clustered together with high bootstrap values at the tips and deeper relationships were also recovered in congruence with morphological and nuclear DNA data. An area of incongruence was also clear between the nodes at the tips and the deeper nodes, with intermediate nodes showing morphologically unexpected or geographically surprising relationships with weak support [12]. Therefore, the phylogenetic results from barcodes were considered mixed. Despite conflict, the barcode data were promising in that a small amount of rRNA data (estimated to be more appropriate for capturing intermediate and deeper nodes [11]) seemed able to dominate the combined data. In other words, although the number of variable characters from the barcodes was far greater than the rRNA, the barcode data did not appear to carry any strong biases, and was able to inform on the tips of the tree without negatively influencing deeper nodes. Even though it is clear that COI is not an optimal gene for deep-level phylogenetics [11] and single genes may not reflect species phylogeny (gene trees may not match species trees, especially at the shallow parts of the tree toward the terminals) owing to independent sorting, introgression or other problems [19,20]), the relevant question is: 'Are COI-generated phylogenetic hypotheses worth reporting?' We show that they are.

## 2. Material and methods

### (a) Dataset

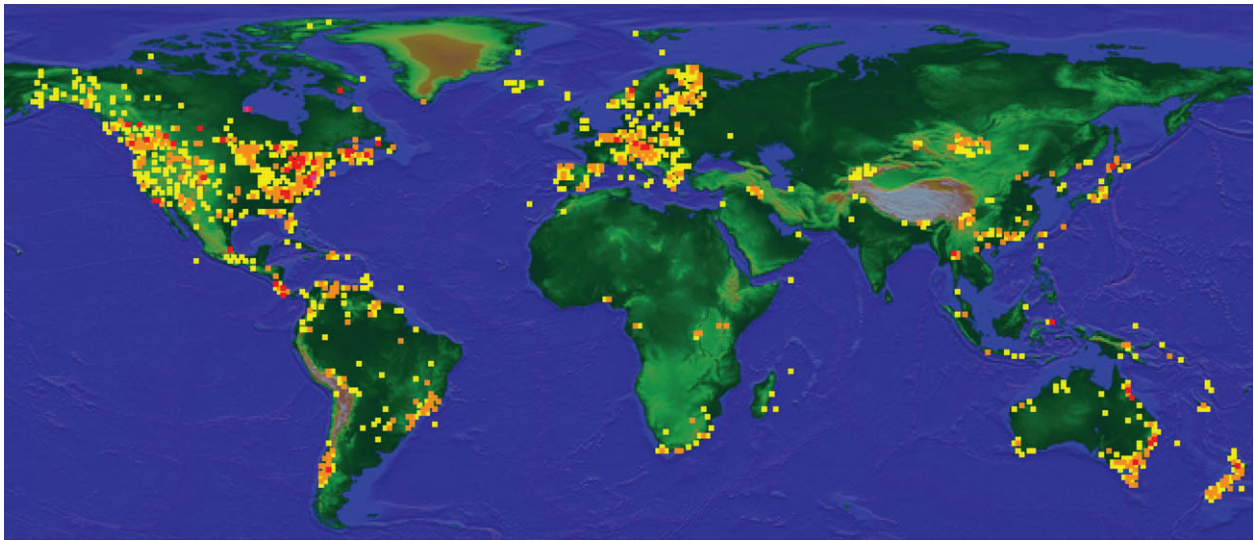
The public records from the BOLD systems were searched for Trichoptera in February 2016, using the following criteria: more than 500 bp, not flagged as errors, contaminants or stop codons. Once the FASTA files were downloaded from the BOLD website, we applied a script that identified and merged all identical haplotypes into a single OTU (operational taxonomic unit, which in this context is equivalent to a unique haplotype). Additional scripts were written to reduce the taxon labels provided by BOLD to include only the species name, a three-letter locality abbreviation indicating 'country' and numerical codes that indicate the number of individuals that possess that identical haplotype. The seven largest countries (by area) received a second two-letter abbreviation for state or province. If individuals with identical haplotypes were collected in different locations, each location is represented in the taxon name. We uploaded all relevant files to a GitHub repository ([https://github.com/pbfrandsen/trichoptera\\_barcode](https://github.com/pbfrandsen/trichoptera_barcode)), including all scripts and detailed sample information. Specimen IDs are recorded according to their labels on the tree and available as an Excel file on GitHub. Specimen information, along with sequences, electropherograms and primer details for each specimen are available in BOLD at the DOI [dx.doi.org/10.5883/DS-TBOL](https://doi.org/10.5883/DS-TBOL) and in GenBank (accession numbers KX291053–KX296688).

### (b) Phylogenetic inference

#### (i) Phylogenetic constraints

Because the COI gene has been shown to be homoplastic for recovering deep-level phylogenetic trees for Trichoptera [11,12], we applied a series of topological constraints to our analysis to concentrate the resolving power of the barcode data toward the tips of the tree. These constraints were generated from a variety





**Figure 1.** Geographical distribution of public Trichoptera records from the BOLD website (2016). Yellow, 1–9; orange, 10–99; red, 100–999 sample records.

of sources, using a variety of criteria. The primary source of constraints came from our most recent phylogeny [21], generated from multiple genes, and over 10 000 nucleotides; nodes from this work that had bootstrap values above 85%, and/or nodes that were supported by two or more independent loci were constrained. Thus, in this case, independent corroboration was deemed more important than bootstrap support, but in reality, these measures are highly correlated [21]. We also used other smaller datasets (cited below), and for these, we set the constraints for nodes supported by posterior probabilities of 100%, because posterior probabilities are often much higher than bootstraps [22]. Additional taxa from Malm and co-workers [23] were added to the whole order constraint tree. Within individual families, we set constraints for Leptoceridae [24], Polycentropodidae [25] and Glossosomatidae [26], and within genera, for *Chimarra* [12]. The constraint tree is available as a Newick file in the GitHub repository.

## (ii) Phylogenetic reconstruction

Once sequences were downloaded and constraints were established, we analysed them with RAXML [1], under the GTR + GAMMA substitution model. First, we estimated a tree representing all OTUs within Trichoptera. The constraint tree described above was used to guide the tree search via the ‘-g’ option in RAXML, which allows the user to specify a multifurcating constraint tree for a subset of the taxa or all taxa in the alignment. Taxa not represented in the constraint tree were then placed into the scaffold phylogeny induced by the constraint. The taxa not present in the constraint tree are therefore placed by the barcode data, but because the constraint tree is fixed, it is probable that new taxa will attach themselves to a taxon in the constraint tree that is relatively closely related. This strategy works best when the constraint tree is densely populated.

Next, in a separate analysis with different input taxa, we generated trees for smaller, monophyletic groups within Trichoptera (usually families). For each of these groups, we downloaded the barcodes from BOLD, including several outgroups. Then we merged duplicate haplotypes and encoded the taxon labels using the same method and scripts that we used for the order-wide tree. We aligned the sequences in MUSCLE [27] and then partitioned each group into four subsets using a site-specific rate model described by Kjer & Honeycutt [28]. The best-known trees were then estimated in RAXML with 1000 rapid bootstraps (using the ‘-f’ option) [1]. Thus, there are two separate analyses: one for the all taxon tree and another for subgroups. In the all taxon tree, if there is a contaminant that is labelled, for example,

as ‘Xiphocentronidae’, but it is in reality an Ecnomidae, it will be placed with the Ecnomidae in the big tree. By contrast, it will appear as an extremely long branch (and obvious red flag) on the Xiphocentronidae tree, because it is not a xiphocentronid [29].

## (iii) Tree presentation

Presenting all the trees generated for this paper in print (even only for unique haplotypes), with a readable font size, would require nearly 160 pages. This demonstrates one of the challenges in presenting the ‘Tree of Life’, even for a moderately sized insect order with incomplete taxon sampling. We solved this problem in two ways: first, all trees were uploaded to the GitHub repository in Newick format. Each can easily be downloaded and viewed with a program such as ‘FigTree’ [30] or ‘Dendroscope’ [31]; second, the larger (all taxon) tree is also available on the iTOL website (<http://itol.embl.de/tree/16011125417288281456757921>).

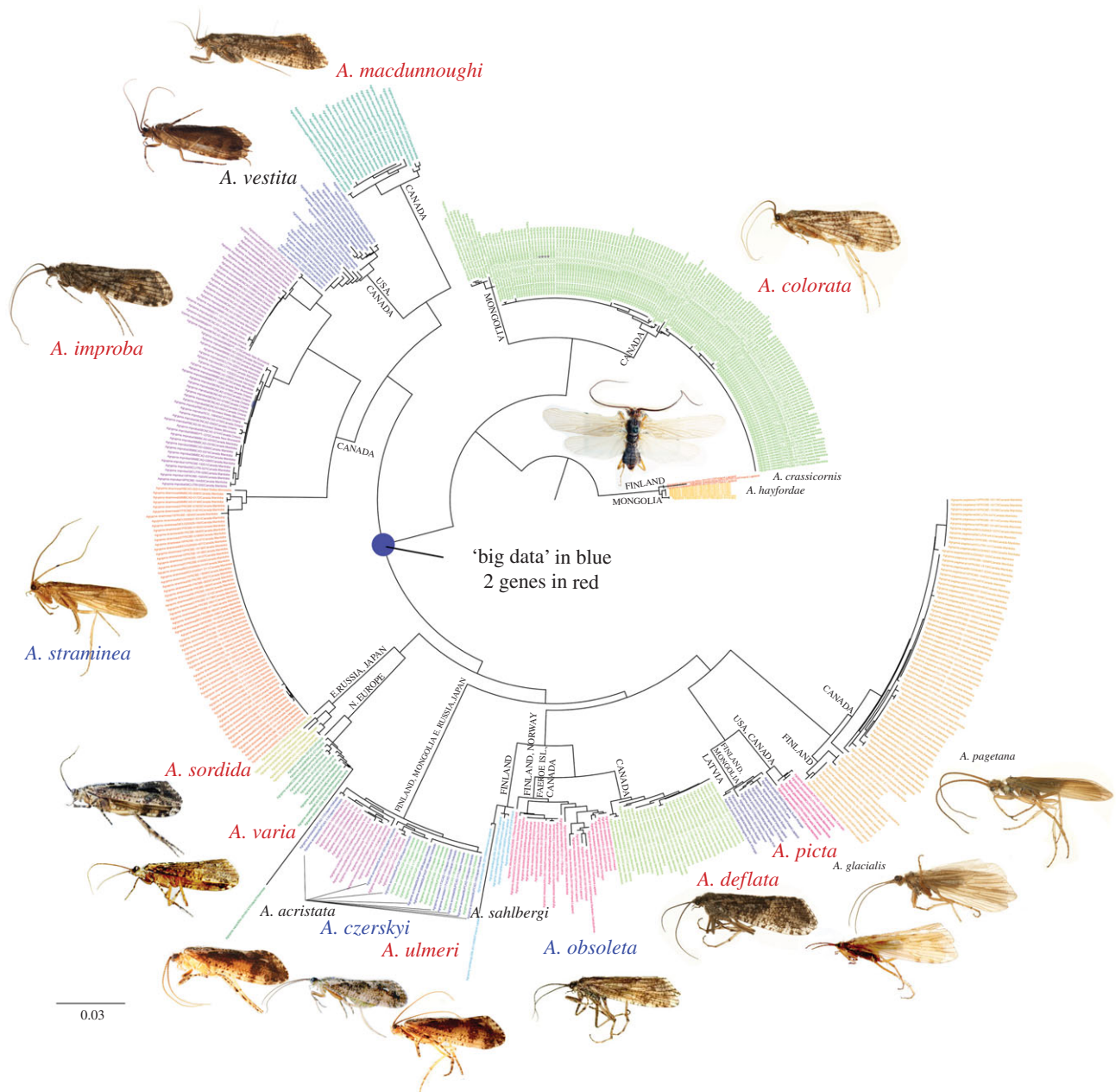
# 3. Results

## (a) The dataset

A total of 49 932 records were recovered from BOLD, of which 38 999 barcodes met our criteria. This dataset includes 5569 ‘BINs’, representing 3280 named species (including 33 subspecies) and 1009 interim names (provisional identifications used by BOLD users, a system often applied by taxonomists as working hypotheses to highlight potentially new or unidentified species), from 484 genera and all 49 Trichoptera families. A ‘BIN’ is a DNA-barcode-based registration system for animals, which represents a cluster of haplotypes, aiming to facilitate revisionary taxonomy [32]. Specimens came from 81 countries, with concentrations visible on the map generated from BOLD (figure 1). Viewing the BOLD taxonomy browser in February 2016 for Trichoptera ([http://www.boldsystems.org/index.php/Taxbrowser\\_Taxonpage?taxid=99](http://www.boldsystems.org/index.php/Taxbrowser_Taxonpage?taxid=99)), we see a concentration of sampling efforts at the Biodiversity Institute of Ontario (BIO, 19 783: 34%), the University of Minnesota Insect Collection (UMSP, 8301: 14%), the National Museum of Natural History, Smithsonian Institution (NMNH, 6,376: 11%) and others (23 733: 41%). However, the number of individuals processed and the number of species added to the database are not tightly correlated because many of the individual samples come from bio-survey efforts, sometimes

The results from our work on *Chimarra* [12] encouraged us to employ DNA barcodes in phylogenetic reconstruction (but with caution). Our phylogeny for the entire order is shown in figure 2. There are two ways to visualize the details of the tree. First, a Newick file of this tree is available for download from GitHub and can be viewed in phylogenetic tree viewing





**Figure 3.** Phylogeny of *Agrypnia*, demonstrating the concept of using barcode data to improve the taxon sampling at the tips of the tree, and using larger datasets as topological backbone constraints. Species names in blue: those represented by hundreds of genes; species in red: those represented by COI barcode and ribosomal DNA; species names in black: those only represented by COI barcode.

software. Second (recommended), the tree is visualized on the iTOL website (<http://itol.embl.de/tree/16011125417288281456757921>). This file will be periodically updated by the Trichoptera Barcode of Life community. The tree shown in figure 2 will be labelled as 'Trichoptera\_barcode\_tree\_28\_Feb\_2016'.

Another resource available for download is individual trees for trichopteran families. Compared with the larger tree, errors in these trees are easier to detect. For the larger analysis (figure 2), RAxML was given a tree with individual haplotypes as a backbone, upon which the RAxML program grafted additional haplotypes where they best fit without changing the order of the branches on the constraint tree. This is ideal, because if a specimen is misidentified, or is a contaminant, then it will attach to the tree by what it is, not what it is named in the database. Most of these misplaced taxa are the result of errors in data submission, or errors in identification (most often females or larvae), or unidentified contamination errors, as suggested by long branches in figure 2. These long branches are intentionally

retained. It will take years to clean up the entire dataset, as it is continually growing and corrections often involve specimen loans and careful examination of the vouchers. We do not think it wise to throw some out without a clear criterion for doing so. In addition, we contend that these errors need to be visualized before the painstaking process of curating the dataset is completed. For example, we noted that some of the samples from far Eastern Russia have barcodes that are identical across families. While re-examining the Excel sheets we used to submit these data, we found a transposition error that resulted in some taxa being mislabelled. These errors are obvious in the tree file and will be corrected in subsequent online revisions.

### (c) Using barcodes for integrative taxonomy

The large Trichoptera barcode phylogeny tree provides an invaluable basis for careful scrutiny of species boundaries and relevant hypotheses. For instance, figure 3 provides an

example of integrating barcode data into a previous phylogenetic hypothesis. Here, morphology, traditional Sanger-based molecular analysis and transcriptome data placed *Agrypnia* within Phryganeidae [21]. We generated a large dataset of Phryganeidae [33] that included three *Agrypnia* species, *A. straminea*, *A. obsoleta* and *A. czerskyi*, covering 892 genes. *Banksiola* and *Oligotricha* were selected as outgroup taxa [33]. We also generated 28S sequences for seven additional *Agrypnia* species, all of which had barcode data. Finally, we added the barcode data to the backbone constraints generated by the larger datasets.

When examining the phylogeny (figure 3) at the species-level, we see many species whose haplotype clusters are distinct, and reflect the names they have been given. Most of the species in figure 3 are preceded by long internodes that reflect genetic isolation that match our conceptual ideas about what species are. However, others are a mixture. The barcode haplotypes for *A. czerskyi*, *A. cristata* and *A. ulmeri* are intercalated among one another (figure 3). This does not necessarily mean that they are not real species. Species are defined by taxonomists using integrative datasets, not barcodes alone. However, figure 3 should encourage a taxonomist to reconsider the evidence for this cluster of near identical barcode haplotypes.

## 4. Discussion

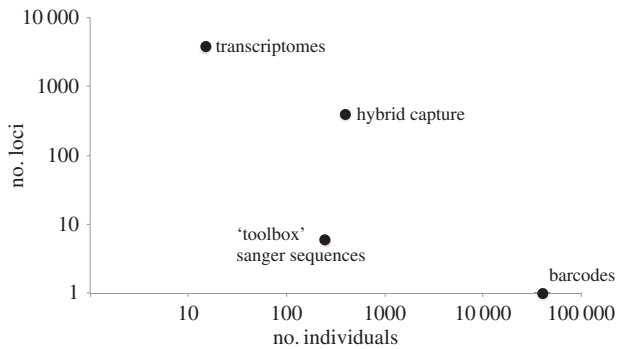
In this paper, we presented the largest Trichoptera phylogeny to date, which was built based on data from multiple genes for deep-level nodes and novel barcode data for terminal tips. This large phylogeny (figure 2) contains a great deal of potentially useful information. For example, taxonomists working with any species in the tree can find hypothesized sister species that are probably good outgroups. Both taxonomic and geographical consistency clearly demonstrate that this is a meaningful tree. In our most recent phylogeny [29], the barcode fragment alone recovered Trichoptera, Annulipalpia, Integripalpia, Brevitentoria and Plenitentoria—virtually the entire backbone of deep Trichoptera phylogeny. Similar results were seen across Trichoptera [34] and in *Chimarra* [12]. However, all these studies noted that intermediate nodes were problematic when assessed by congruence to morphology and/or other molecular data. A thorough discussion of ‘why?’ is beyond the scope of this paper, and it is difficult to separate nodes we ‘like’ from those we do not. Still, the tree is likely to contain two sources of inaccuracy. First, like all phylogenetic hypotheses, our constraint tree is likely to contain errors. Constraints were built from much larger, published datasets and can be modified as our future understanding improves. Second, inaccuracies can arise from the limitations of the COI data itself. The barcode is, of course, only a short fragment of a single, rapidly evolving locus. We suspect that the analysis of any short fragment of DNA will be subject to stochastic error. In deep parts of the tree, we suspect that the barcode fragment will be saturated [11] (although it still recovers many deep nodes on its own). However, some of these problems are alleviated by the constraints and have little effect on intermediate nodes. Confounding signal from incomplete lineage sorting (at the shallow nodes) or historical hybridization and introgression (at the deeper nodes) is a possibility. We note that these are not ‘errors’, but rather, a reflection of true biological processes when they are accurately estimated. However, they are in conflict with the species tree. For many groups of caddisflies, however,

the barcode data provide the only phylogenetic hypothesis we have. How ‘good’ the tree is, especially at the shallow nodes, can only be discovered with additional data. We find these phylogenetic hypotheses to be useful for circumscribing potential species, for inferring a first hypothesis of relationships among species, for identifying misidentified taxa, and for associating unidentified life-history stages with their described adult stages. We are committed to continually updating our phylogeny. As we add more genes, we can address congruence, decrease stochastic errors and increase node support. As we add more taxa, even with a single gene, other problems owing to long-branch effects, for example, will be reduced. Here, we present our current best hypothesis. We would recommend, given the nature of the data, that nodes with low bootstrap values should be ignored or discarded. Contamination, misidentification and misclassification in public databases, including BOLD, are genuine problems that need to be addressed. Here, we provide two means to detect them. They will appear as long-branch taxa in the family-level trees on GitHub (which were analysed together according to the names assigned to them, which are sometimes wrong). A search through the whole taxon tree (figure 2) for those same haplotypes will offer a hypothesis of what they really are, as you might find members of different suborders appearing to have identical, or nearly identical haplotypes. This is because in the big tree, all taxa are analysed together, and mislabelled taxa will find their proper place in the phylogeny, despite the error in the label. Either case should direct researchers toward the voucher specimens for either correction or reclassification.

There have been attempts to automate tree building, tapping into public databases to produce large summary phylogenies. For example, researchers have explored the possibility of producing trees from GenBank of up to 8000 taxa [35,36]. However, these trees were subject to the errors in the database (as indicated above) and the limitations of the few genes that dominate in GenBank. We find the prospect for non-specialists to produce an automated phylogeny at the push of a button to be exciting, but such a phylogeny is only as good as the care that went into producing the database and in the rigour of the phylogenetic analysis. Our recommendation is a distinct alternative, in that we think specialists should evaluate phylogenies from reliable sources, to update current hypotheses based on constraints, using specified, transparent criteria. These constraints can then be used as a backbone on which to hang other data such as barcodes. Barcodes are ideal for what they were designed to do: distinguish species, and even populations, from one another. However, phylograms with extremely long terminal branches relative to short intermediate and deep internodes are problematic for phylogenies [11], and the BOLD tree option, which uses Neighbour-Joining [37], was not designed to produce phylogenies (although it is excellent for finding the closest match across the datasets, which is what it was designed to do). We decided which constraints were reasonable and justified this based on our expertise in Trichoptera. The advantage of an expertly curated phylogeny available on the World Wide Web is that an up-to-date phylogeny will be available beyond the print version of this paper, and for as long as the authors are able.

## (a) A plan for future phylogenetic work

In Trichoptera phylogenetics, there are four sources of molecular data that are quite different in scale, each targeted at



**Figure 4.** Comparison of dataset sizes in terms of number of taxa and number of loci.

different levels of divergence: transcriptomes, targeted enrichment sequences [38,39], PCR-based Sanger nuclear sequences and barcodes. We have generated a data matrix from six of the standard 'toolbox' genes, comprising approximately 12 000 combined nucleotides for approximately 250 taxa [11,40,41]. Others have also made significant contributions to the higher-level phylogeny of Trichoptera and their work also contributed to the constraint tree [23–26]. Transcriptome data are being collected through the 1000 Insect Transcriptome Evolution project (1KITE, [www.1KITE.org](http://www.1KITE.org)), and these data dwarf the previous dataset, with 3800 nuclear genes for 60 trichopteran taxa (figure 4). However, transcriptome sequencing requires the collection of fresh, specially preserved specimens, which makes the inclusion of rare taxa difficult. To fill these taxonomic gaps, we have been generating data with hybrid capture techniques [38] and have captured 900 loci for specimens representing 250 Trichoptera genera. A strategy for using these large datasets to provide a backbone phylogeny was briefly discussed elsewhere [33]. It should be possible to generate transcriptomes from every family, hybrid capture data from every obtainable genus, rRNA (ideally the 28S, D2) from thousands of species, and barcode data for most species. Our strategy is to use large, multigene datasets to generate a backbone tree, and then use barcode data to fill in the leaves of the tree. It is our intention to continually update this tree as new constraints are discovered through our future analyses of large datasets.

### (b) DNA barcodes and integrative taxonomy

Repeating the process illustrated in figure 3 (for *Agrypnia*) for the other 615 extant genera would be a difficult task, but not impossible to imagine. As datasets grow, particularly from high-throughput sequencing, constraints will improve. Every population does not require a genome to place it in a phylogenetic context. What is missing from our example (figure 3) is taxonomic insight. We believe this would greatly improve the final result and we expect that specialists will contribute summary works on their genera, adding data to the barcodes, as was done with *Chimarra* [12], *Neophylax* [42] and Drusinae [43,44] using two or more genes and morphological expertise.

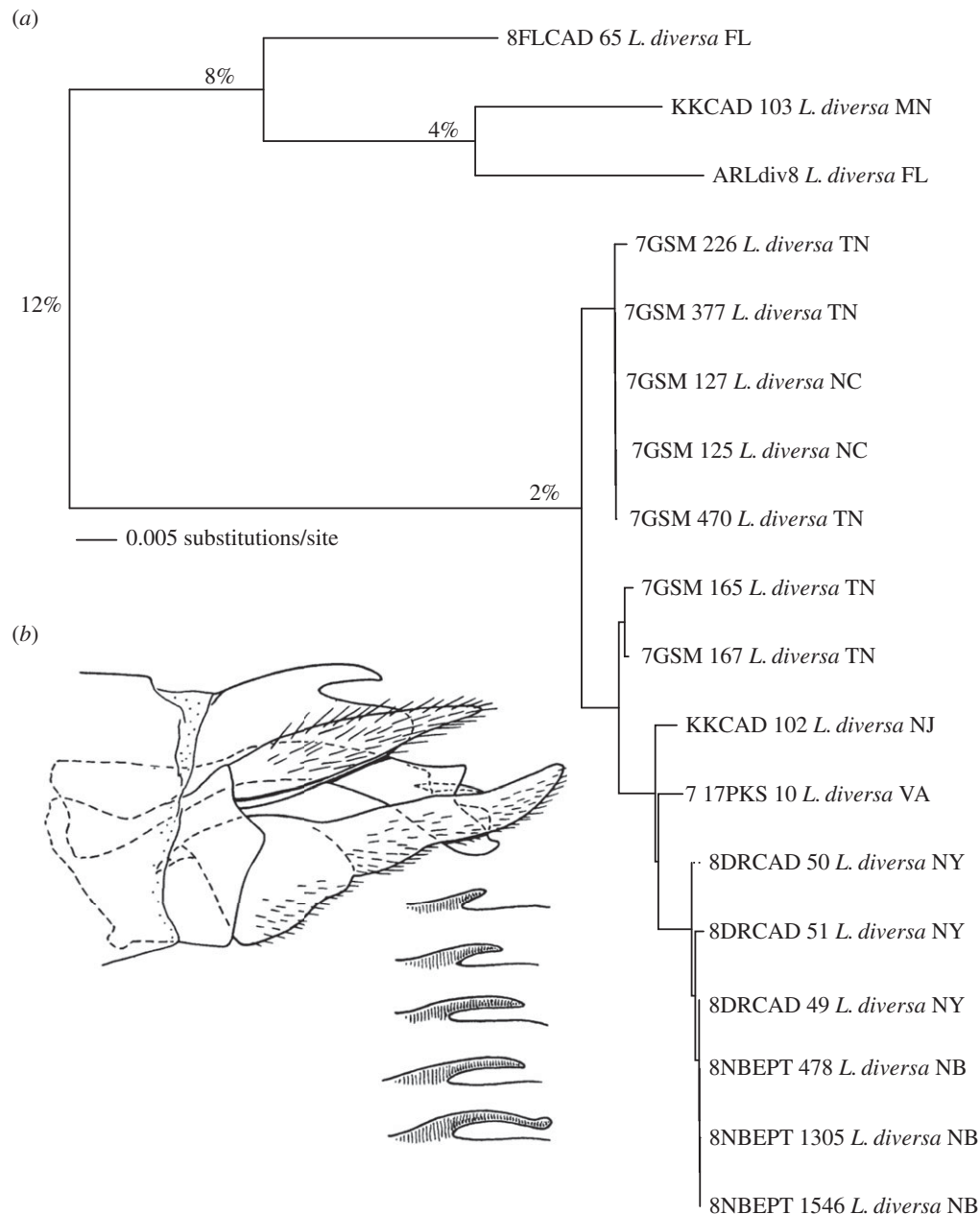
An example of how barcodes can inform descriptive taxonomy is shown in figure 5. There is a hint that morphological variation was recognized from the description, in the species epithet of *Lype diversa* Banks (Psychomyiidae). Ross [45] illustrated variation in the genitalia of this species

(figure 5b). Still, it has remained a single species since its description in 1914. The variation in branch-lengths in the barcode phylogram points toward hidden biodiversity. Using a 2% threshold, a number that is often correlated with species-level diversity, there are five species. However, biological species in reality are not based on pairwise haplotype divergence. Useful algorithms can be constructed that estimate mean differentiation within circumscribed lineages, and these algorithms are useful for estimating diversity in biological sample assemblages, as well as the probable number of species clusters in aggregate. But on the individual level, species should not be defined on expedient algorithms. We recommend that haplotype clusters (as shown in figures 2, 3 and 5) can be explored by taxonomists in search of species hypotheses. For example, if KKCAD-103 and ARLdiv8 shared a straight stub on the 10th tergite that looked like the top illustration in the inset to figure 5b, and 8FLCAD-65's spine looked like the spine at the bottom, while all the rest had curved, but not elongate spines, lacking the knob on the end, then with confirmation of these patterns from museum specimens, and consideration of geographical patterns, a taxonomist could confidently describe three species from figure 5, secure in the application of a species definition that matches our concepts of speciation. However, to revise the *L. diversa* species complex, additional specimens should be examined across the range of the species, the morphological characters should be described and analysed in detail, the type species and any synonyms should be examined to determine which available names apply to which morphotypes, the new species and their variation should be carefully illustrated or photographed, new names should be proposed as needed, and specimens should be vouchered, type specimens designated and all specimens deposited in a publicly accessible museum. These requirements are difficult in today's funding climate and we are dismayed at the continued erosion of support for taxonomy.

The patterns illustrated in figure 5 are not uncommon. The classic paper by Hebert and co-workers [46] is a prime example of cryptic biodiversity discovery. Years of observations on larval diet and morphology of *Astraptus fulgurator* (Lepidoptera, Hesperidae) did not lead to serious species descriptions (ignoring the precedent of Brower [47], which we would not follow as an example), because the adults were indistinguishable, and in this case, it was difficult to decide if the observed variation was the result of environmental factors. It is the task of taxonomists to decide if the variation they observe is intraspecific or interspecific. In the above given example, Brower proposed to accept formally 3 to 7 of the 10 barcode clusters at species rank [48]. Barcode data provide corroborative evidence, but cannot stand alone without reciprocal illumination from other genes and/or morphology. A similar example was shown by Harvey and co-workers [49], who found that distinct haplotype clusters precisely matched fixed variation in larval head morphology of the *Diplectrona modesta* species complex, a widely distributed caddisfly across eastern North America (see their fig. 2).

We have witnessed an alarming decline in the support for museums and organismal taxonomy. The Trichoptera barcode database owes much of its utility to the tireless efforts of specialist taxonomists collecting and identifying specimens. Without these identifications, the database would show only a large collection of DNA haplotypes without associated species names. Species names have been a baseline for biological work for





**Figure 5.** Example of corroborative species illumination. (a) Neighbour-joining phylogram from BOLD of *Lype diversa* (Psychomyiidae). Taxa are labelled by an abbreviation of the specimen ID, followed by the species name and then state or province abbreviation. Numerals on the internodes (in per cent) represent mean pairwise distances between the nodes directly to the right. (b) Illustration from [45], ‘showing variations of the dorsal horn of the 10th tergite’.

more than 250 years, and should not be abandoned for BINs or MOTUs (Molecular Operational Taxonomic Units). Earlier work on Trichoptera has shown that the number of haplotype clusters in an environmental sample is tightly correlated with the number of species [50], so that BINs in aggregate can provide information that can be decoupled from taxonomy. For some applications, such as biodiversity inventories, this correlation would translate into fast, efficient and accurate biodiversity assessments and community ecology studies. However, abandoning the process of species description altogether for the expediency of single-gene based BINs would be a mistake. Barcode data can, however, assist in corroborating morphological species boundaries, as shown in figure 5, and demonstrated by e.g. the work of Flint & Kjer [42] and Previšić and co-workers [43]. As the barcode database grows at its own rapid pace, it will be increasingly valuable in distinguishing between intra-specific and interspecific variations [51]. The two systems (generation of barcode libraries and the description of species)

can advance independently, but should not be entirely decoupled. Haplotype clusters can be algorithmically defined, and should be called ‘BINs’ (where obtained from BOLD’s BIN System) or ‘MOTUs’ from other molecular analysis algorithms. Congruent genetic clusters from multiple genes or whole genomes are probably distinct species, but we should still refrain from calling them species until they have been formally described by taxonomists. An example of the merging of DNA taxonomy with biologically sound species concepts was presented for Chinese Hydropsychidae; these authors made recommendations for associating larvae with adults that coupled congruence across multiple genes with morphological hypotheses of species [20]. These kinds of associations can be used to describe the morphology of the larvae to aid the use of these species in water quality assessment [50,52,53]. Similarly, barcode evidence, together with morphology, has been employed in differentiating cryptic diversity and defining new species of South American and European caddisflies [43,54,55]. Thus, we find

that barcodes provide a valuable tool in answering a wide variety of ecological and taxonomic questions. However, without support for taxonomy, where expertise is rapidly declining, we condemn ourselves to see 'biodiversity' as a collection of plastic tubes, named with alpha-numeric codes that are divorced from biological context.

**Authors' contributions.** X.Z., P.B.F., R.W.H. and K.M.K. conceived of the work and wrote the paper. X.Z. led the Trichoptera Barcode of Life initiative and oversaw most of the data collection. P.B.F. wrote the scripts and analysed the data. K.M.K. received specimens from collaborators, prepared them and sent them to X.Z. R.J.B., R.W.H. and O.S.F. curated most of the specimens (in terms of species) and facilitated sampling in their respective museums. H.M. provided an enormous number of specimens. M.A.M. identified the public BOLD data, prepared it for analysis and submitted it to GenBank. All other authors (those listed in alphabetical order) provided a significant number of specimens or sequences and helped with editing the manuscript and addressing software issues. All authors gave final approval for publication.

**Competing interests.** We have no competing interests.

**Funding.** The majority of barcode sequences were generated at the Centre for Biodiversity Genomics, through the iBOL initiative and the Trichoptera Barcode of Life project, funded by the Ontario Ministry of Research and Innovation and by the government of Canada through Genome Canada. X.Z. contributed Chinese materials and data through supports from the Ministry of Science and Technology of the People's Republic of China through the National High-tech Research and Development Project (863) of China [2012AA021601] and the National Science and Technology Support Program of China [2012BAK11B06-4]. K.M.K. thanks the Schlinger endowment for funding, and US National Science Foundation grants DEB-0316504 and DEB-0816865. Support for work associated with the University of Minnesota Insect Collection came from Minnesota Experiment Station Projects 017-017 and 017-029 and the German Academy of Sciences Leopoldina Postdoctoral Fellowship awarded to S.U.P. (BMBF-LPD 9901/8-169). The collection and processing of specimens were partly funded as components of the

'Barcoding Fauna Bavarica' project by the Bavarian Ministry of Science, Research and Art (Bayerisches Staatsministerium für Wissenschaft, Forschung und Kunst, Munich, Germany), the 'German Barcode of Life' project by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, 01LI1101B–01LI1501B). The Finnish Trichoptera barcoding project is funded by the Ministry of the Environment of Finland, Finnish Barcode of Life Project, and Academy of Finland. J.R. would like to acknowledge the University of Tennessee-Knoxville Ecology and Evolutionary Biology. Collecting and sequencing of specimens from D.R. Congo was supported by the Field Museum Council on Africa awarded to S.U.P. and F.N.M. Sequencing of Drusinae material was partly funded by Austrian Science Fund (FWF) project P23687-B17 awarded to J. Waringer, W. Graf and S.U.P. The EPA Victoria and the Murray-Darling Basin Authority of Australia funded sequencing of some Australian specimens. R.S.C. would like to acknowledge the EPA Victoria in Australia for supporting submission of specimens for the BOLD database and the Department of Environment, Land, Water and Planning for funding. A.S. was financially supported by the Klaus Tschira Foundation. Research funding in New Zealand was provided to I.D.H. by the New Zealand Foundation for Research, Science and Technology (UOWX0501 and UOWX0505) and the Terrestrial and Freshwater Biodiversity Investment System (TFBIS) Programme Numbers 246 and 290. Spanish specimens provided by C.Z.-M. and N.B. came from the projects 70/04/P/E financed by the Spanish Agency for International Cooperation, and the GUADALMED (HID98-0323-C05 and REN2001-3438-C07) and the RICHABUN (CGL2007-60163/BOS) projects funded by the Spanish Ministry of Education and Science, and the project ref. 039/2007 financed by the Autonomous Organism of National Parks (O.A.P.N.) of the Spanish Ministry of Environment. Barcoding work by S.C. and J.C.M. in the Mongolian Aquatic Insect Survey was supported by the US National Science Foundation (DEB-0206674 and DEB0743732). Finally, many collaborators contributed samples using personal funds.

**Acknowledgements.** Thanks to Paul Hebert for his encouragement, funding and vision. We thank Marta Sáinz-Bariáin for samplings from Sierra Nevada, Spain and Dave Ruiters for his substantial contributions to the database.

## References

1. Stamatakis A. 2014 RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
2. Price MN, Dehal PS, Arkin AP. 2010 FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490. (doi:10.1371/journal.pone.0009490)
3. Hinchliff CE *et al.* 2015 Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl Acad. Sci. USA* **112**, 12 764–12 769. (doi:10.1073/pnas.1423041112)
4. Misof B *et al.* 2014 Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767. (doi:10.1126/science.1257570)
5. Jarvis ED *et al.* 2014 Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331. (doi:10.1126/science.1253451)
6. Wickett NJ *et al.* 2014 Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl Acad. Sci. USA* **111**, E4859–E4868. (doi:10.1073/pnas.1323926111)
7. Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218)
8. Sperling FA, Anderson GS, Hickey DA. 1994 A DNA-based approach to the identification of insect species used for postmortem interval estimation. *J. Forensic Sci.* **39**, 418–427. (doi:10.1520/JFS13613J)
9. Castresana J. 2001 Cytochrome *b* phylogeny and the taxonomy of great apes and mammals. *Mol. Biol. Evol.* **18**, 465–471. (doi:10.1093/oxfordjournals.molbev.a003825)
10. Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74. (doi:10.1016/S0169-5347(02)00041-1)
11. Kjer KM, Blahnik RJ, Holzenthal RW. 2001 Phylogeny of Trichoptera (Caddisflies): characterization of signal and noise within multiple datasets. *Syst. Biol.* **50**, 781–816. (doi:10.1080/106351501753462812)
12. Kjer KM, Zhou X, Frandsen PB, Thomas JA, Blahnik RJ. 2014 Moving toward species-level phylogeny using ribosomal DNA and COI barcodes: an example from the diverse caddisfly genus *Chimarra* (Trichoptera: Philopotamidae). *Arthropod Syst. Phylogeny* **72**, 345–354.
13. Holzenthal RW, Blahnik RJ, Prather AL, Kjer KM. 2007 Order Trichoptera Kirby, 1813 (Insecta), caddisflies. *Zootaxa* **1668**, 639–698.
14. Holzenthal RW, Thomson RE, Rios-Touma B. 2015 Order Trichoptera. In *Ecology and general biology, Vol 1: Thorp and Cowich's freshwater invertebrates*, (eds JH Thorp, DC Rogers), pp. 965–1002, 4th edn. New York, NY: Academic Press.
15. Rosenberg DM, Resh VH. 1993 *Freshwater biomonitoring and benthic macroinvertebrates*, p. 488. New York, NY: Chapman & Hall.
16. Kristensen NP. 1999 Phylogeny of endopterygote insects, the most successful lineage of living organisms. *Eur. J. Entomol.* **96**, 237–253.
17. Willis JC, Yule GU. 1922 Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature* **109**, 177–179. (doi:10.1038/109177a0)
18. Morse JC. 2013 Trichoptera World Checklist. See <http://entweb.clemson.edu/database/trichopt/index.htm>.
19. Avise JC. 1989 Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* **43**, 1192–1208. (doi:10.2307/2409356)
20. Zhou X, Kjer KM, Morse JC. 2007 Associating larvae and adults of Chinese Hydropsychidae caddisflies (Insecta:Trichoptera) using DNA sequences. *J. N. Am. Benthol. Soc.* **26**, 719–742. (doi:10.1899/06-089.1)
21. Kjer KM, Thomas J, Zhou X, Frandsen PB, Scott E, Holzenthal RW. 2016 Progress on the phylogeny of caddisflies (Trichoptera). In *Proc. of the 14th Int.*

- Sym. on Trichoptera, July 2–7, 2012, Vladivostok, Russia* (eds TS Vshivkova, JC Morse), p. 216–224. Auckland, NZ: Magnolia Press.
22. Lewis PO, Holder MT, Holsinger KE. 2005 Polytomies and Bayesian phylogenetic inference. *Syst. Biol.* **54**, 241–253. (doi:10.1080/10635150590924208)
  23. Malm T, Johanson KA, Wahlberg N. 2013 The evolutionary history of Trichoptera (Insecta): a case of successful adaptation to life in freshwater. *Syst. Entomol.* **38**, 459–473. (doi:10.1111/syen.12016)
  24. Malm T, Johanson KA. 2011 A new classification of the long-horned caddisflies (Trichoptera: Leptoceridae) based on molecular data. *BMC Evol. Biol.* **11**, 10. (doi:10.1186/1471-2148-11-10)
  25. Johanson KA, Malm T, Espeland M, Weingartner E. 2012 Phylogeny of the Polycentropodidae (Insecta: Trichoptera) based on protein-coding genes reveal non-monophyletic genera. *Mol. Phylogenet. Evol.* **65**, 126–135. (doi:10.1016/j.ympev.2012.05.029)
  26. Robertson DR, Holzenthal RW. 2013 Revision and phylogeny of the caddisfly subfamily Protoptilinae (Trichoptera: Glossosomatidae) inferred from adult morphology and mitochondrial DNA. *Zootaxa* **3723**, 1–99. (doi:10.11646/zootaxa.3723.1.1)
  27. Edgar C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797. (doi:10.1093/nar/gkh340)
  28. Kjer KM, Honeycutt RL. 2007 Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol. Biol.* **7**, 8. (doi:10.1186/1471-2148-7-8)
  29. Frandsen PB, Flint Jr OS, Zhou X, Kjer KM. 2016 A proposal for using barcode data to fill out the leaves of the Trichoptera tree of life. In *Proc. of the 14th Int. Symp. on Trichoptera, July 2–7, 2012, Vladivostok, Russia* (eds V TS, M JC), pp. 161–167. Auckland, NZ: Magnolia Press.
  30. Rambaut A. 2012 Figtree v1.4.0, a graphical viewer of phylogenetic trees. See <http://tree.bio.ed.ac.uk/software/figtree>.
  31. Huson DH, Scornavacca C. 2012 Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067. (doi:10.1093/sysbio/sys062)
  32. Ratnasingham S, Hebert P. 2013 A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE* **8**, e66213. (doi:10.1371/journal.pone.0066213)
  33. Frandsen PB. 2015 Large datasets and Trichoptera phylogenetics. PhD dissertation, Rutgers University, NJ.
  34. Hogg ID, Smith BJ, Banks JC, Dewaard JR, Hebert PDN. 2009 Testing use of mitochondrial COI sequences for the identification and phylogenetic analysis of New Zealand caddisflies (Trichoptera). *New Zealand J. Mar. Freshw. Res.* **43**, 1137–1146. (doi:10.1080/00288330.2009.9626536)
  35. Bocak L, Barton C, Crampton-Platt A, Chesters D, Ahrens D, Vogler AP. 2014 Building the Coleoptera tree-of-life for > 8000 species: composition of public DNA data and fit with Linnaean classification. *Syst. Entomol.* **39**, 97–110. (doi:10.1111/syen.12037)
  36. Peters RS, Meyer B, Krogmann L, Börner J, Meusemann K, Schütte K, Niehuis O, Misof B. 2011 The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biol.* **9**, 55. (doi:10.1186/1741-7007-9-55)
  37. Saitou N, Nei M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
  38. Lemmon AR, Emme SA, Lemmon EM. 2012 Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* **61**, 727–744. (doi:10.1093/sysbio/sys049)
  39. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012 Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* **61**, 717–726. (doi:10.1093/sysbio/sys004)
  40. Kjer KM, Blahnik RJ, Holzenthal RW. 2002 Phylogeny of caddisflies (Insecta, Trichoptera). *Zool. Scrip.* **31**, 83–91. (doi:10.1046/j.0300-3256.2001.00079.x)
  41. Holzenthal RW, Blahnik RJ, Kjer KM, Prather AP. 2007 An update on the phylogeny of caddisflies (Trichoptera). In *Proc. of the 12th Int. Symp. on Trichoptera* (eds J Bueno-Soria, R Barba-Alvarez, B Armitage), pp. 143–153. Columbus, OH: The Caddis Press.
  42. Flint Jr OS, Kjer KM. 2011 A new species of *Neophylax* from Northern Virginia, USA (Trichoptera: Uenoidae). *Proc. Entomol. Soc. Wash.* **113**, 7–13. (doi:10.4289/0013-8797.113.1.7)
  43. Previšić A, Graf W, Vitecek S, Kućinić M, Bálint M, Keresztes L, Pauls SU, Waringer J. 2014 Cryptic diversity of caddisflies in the Balkans: the curious case of *Ecdisopteryx* species (Trichoptera: Limnephilidae). *Arthropod Syst. Phylogeny* **72**, 309–329.
  44. Vitecek S, Graf W, Previšić A, Kućinić M, Oláh J, Bálint M, Keresztes L, Pauls SU, Waringer J. 2015 A hairy case: the evolution of filtering carnivorous Drusinae (Limnephilidae, Trichoptera). *Mol. Phylogenet. Evol.* **93**, 249–260. (doi:10.1016/j.ympev.2015.07.019)
  45. Ross HH. 1944 The caddisflies or Trichoptera of Illinois. *Bull. Illinois Nat. Hist. Surv.* **23**, 1–326.
  46. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004 Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl Acad. Sci. USA* **101**, 14 812–14 817. (doi:10.1073/pnas.0406166101)
  47. Brower AVZ. 2010 Alleviating the taxonomic impediment of DNA barcoding and setting a bad precedent: names for ten species of '*Astraptes fulgerator*' (Lepidoptera: Hesperidae: Eudaminae) with DNA-based diagnoses. *Syst. Biodivers.* **8**, 485–491. (doi:10.1080/14772000.2010.534512)
  48. Brower AVZ. 2006 Problems with DNA barcodes for species delimitation: 'Ten species' of *Astraptes fulgerator* reassessed (Lepidoptera: Hesperidae). *Syst. Biodivers.* **4**, 127–132. (doi:10.1017/S147720000500191X)
  49. Harvey LE, Geraci CJ, Robinson JL, Morse JC, Kjer KM, Zhou X. 2012 Diversity of mitochondrial and larval morphology characters in the genus *Diplectrona* (Trichoptera: Hydropsychidae) in the eastern United States. *Terr. Arthropod Rev.* **5**, 191–211. (doi:10.1163/187498312x649512)
  50. Zhou X, Adamowicz SJ, Jacobus LM, DeWalt RE, Hebert PDN. 2009 Towards a comprehensive barcode library for arctic life—Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. *Front. Zool.* **6**, 30. (doi:10.1186/1742-9994-6-30)
  51. Bergsten J *et al.* 2012 The effect of geographical scale of sampling on DNA barcoding. *Syst. Biol.* **61**, 851–869. (doi:10.1093/sysbio/sys037)
  52. Ruiter DE, Boyle EE, Zhou X. 2013 DNA barcoding facilitates associations and diagnoses for Trichoptera larvae of the Churchill (Manitoba, Canada) area. *BMC Ecol.* **13**, 5. (doi:10.1186/1472-6785-13-5)
  53. Graf W, Lubini V, Pauls S. 2005 Larval description of *Drusus muelleri* McLachlan, 1868 (Trichoptera: Limnephilidae) with some notes on its ecology and systematic position within the genus *Drusus*. *Ann. Limnol.* **41**, 93–98. (doi:10.1051/limn/2005012)
  54. Pauls SU, Blahnik RJ, Zhou X, Wardwell CT, Holzenthal RW. 2010 DNA barcode data confirm new species and reveal cryptic diversity in Chilean *Smicridea* (*Smicridea*) (Trichoptera: Hydropsychidae). *J. N. Am. Benthol. Soc.* **29**, 1058–1074. (doi:10.1899/09-108.1)
  55. Múrria C, Zamora-Muñoz C, Bonada N, Ribera C, Prat N. 2010 Genetic and morphological approaches to the problematic presence of three *Hydropsyche* species of the *pellucidula* group (Trichoptera: Hydropsychidae) in the westernmost Mediterranean Basin. *Aquat. Insects* **32**, 85–98. (doi:10.1080/01650424.2010.482939)